UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Statistics on the evaluation of education
June 2016, Vol. 9, No. 1, pp. 51–72

# An application of plausible values to the standardized test scoring through simulation[1]

## Una aplicación de valores plausibles a la calificación de pruebas estandarizadas vía simulación

Michel Felipe Córdoba Perozo[a]
mcordoba@contratista.icfes.gov.co

### Abstract

The use of plausible values in large scale standardized tests plays the imputation role required when frame of reference is very large and it's not possible for each person to approach the totality of items put into production. In that case it's turned to the definition of a design by blocks that ensures an appropriate quote of evaluated individuals for each item for the framework to be addressed with sufficiency by the whole population under study. The imputation method consist in finding the distribution a posteriori of the latent feature associated to the individual skill, by weighting the distribution that induces the item response theory model and a latent regression associated with some variables measured in the individual. This paper shows an example of simulation where you can easily see the advantages offered by the method in the aggregated results under this scheme of particular application.

***Keywords*:** plausible values, data imputation, rating equalization, latent feature, standardized tests..

### Resumen

El uso de los valores plausibles en evaluaciones estandarizadas de gran escala desempeña el papel de imputación que se requiere cuando el marco de referencia es muy grande y cada individuo no aborda la totalidad de los ítems puestos en producción. En dicho caso se recurre a la definición de un diseño por bloques que garantice una cuota adecuada de individuos evaluados por cada ítem para que el marco de referencia sea abordado con suficiencia por toda la población objeto de estudio. En general, el método de imputación consiste en encontrar la distribución

[1]Córdoba, M. F. (2016) An application of plausible values to the standardized test scoring through simulation. *Comunicaciones en Estadística*, **9**(1), 51-72.
[a]Msc. Statistician. Subdivision of Statistics, Icfes, Colombia

*a posteriori* del rasgo latente que es asociado a la habilidad del individuo, mediante la ponderación de la distribución que induce el modelo de teoría de respuesta al ítem y una regresión latente asociada a algunas variables medidas en el individuo. El presente artículo muestra un ejemplo de simulación, donde se pueden observar de manera sencilla las bondades que brinda el método en los resultados agregados bajo este esquema de aplicación particular.

***Palabras clave*:** valores plausibles, imputación de datos, rasgo latente, pruebas estandarizadas .

# 1 Introduction

In large scale evaluations the frame of reference that defines a test can be too broad. This implies the number of questions to be tackled for the whole population being so big that it makes difficult that each approaches each of them. This difficulty stimulates the use of a design that randomly assigns each individual only a sample of the items so that one of them is approached by an adequate number of individuals.

As each question is not answered by all individuals in the population, design induces an error product of lost or missing observations in the estimation of statistics that define the population's average skill. This problem can be addressed by implementing any imputation technique, not on the responses of the items but the score of individuals for aggregation purposes. The best-known technique is one of plausible values, which uses the multiple imputation methods.

This article illustrates by a small simulation exercise the use and usefulness of this technique: The second section describes the theory that ranges from adjusting the parameters of the model item response to the multiple imputations that generates plausible values. In the third section, we simulate a population of individuals where the aggregate results are calculated in three different scenarios to visualize the rise of bias and the decline of observed variability when each individual fails to present a part of the test. Finally, the solution of this problem is related when plausible values are used.

# 2 Theoretical framework

## 2.1 Usual one-dimensional models of response to item theory

Item response theory models (IRT) are a special case of generalized linear models (GLM). These establish a relation between the answers to a set of items of an individual whom a test is applied and his latent feature, measured over some defined scale, known by many authors as ability. In the model estimation, the ability is

called IRT score. In a specific way, the probability that a certain individual answer correctly to an item is assumed as a function of $\theta$, the symbol used to denote the characteristic to be measured. In this section, the most general features of models that can be usually found in practice are described (Hulin et al. 1983).

### 2.1.1 The Rasch or one parameter logistic model (1PL)

Rasch model is the simplest of the family of one-dimensional models. This model uses only one parameter to characterize each item and one parameter to characterize each person. For individual i with latent ability $\theta_i$ assumed as one-dimensional, the probability of obtaining a right answer to the item $j$, it is, $P(Y_{ij} = 1|\theta_i)$ with $Y_{ij}$ a binary variable that takes the value 1 if individual i answer correctly to the item $j$, is given by following equation:

$$P(Y_{ij} = 1|\theta_i) = P_j(\theta_i) = \frac{1}{1 + \exp[-K(\theta_i - b_j)]} \tag{1}$$

Where bj is the parameter that defines the difficulty of item: when bj increases the probability of answering correctly the item j decreases and K is a constant that defines the scale, usually defined as 1.702. It's important to mark that when $\theta_i$ is a one-dimensional parameter, is a non-linear scaling of statistics $\tau$: the score obtained by classical theory.

In this model, the probability of answering correctly any item tends to zero for small values of $\theta_i$. This implies that individuals with small values of $\theta_i$ won't have the opportunity to correctly answer an item with high or moderate difficulty, it is, according to this model, these individuals won't have the possibility of guessing the right answer. Another important feature of this model is that if difficulty were the same for all, change in the probability of guessing right according to $\theta$ would be the same. This implies that all considered items have the same discrimination (Hulin et al. 1983).

### 2.1.2 Two parameters logistic model (2PL)

This model adds a second parameter that allows more flexibility in adjusting the probability of correctly answering a set of items when these have different discrimination. Expression is given as following:

$$P(Y_{ij} = 1|\theta_i) = P_j(\theta_i) = \frac{1}{1 + \exp[-a_j(\theta_i - b_j)]} \tag{2}$$

Where $b_j$ is the difficulty of item j as in the previous case and $a_j$ the discrimination. This last parameter controls the exchange rate of probability of answering the item correctly in a vicinity of $\theta$ around the value $b_j$. In this case, the probability of answering correctly any item due to chance tends to zero.

### 2.1.3 Three parameters logistic model (3PL)

Besides the two parameters by item that define the 2PL model, the 3PL model incorporate a third parameter in the equation called the chance parameter or casual success. Its functional way is given by the following expression:

$$P(Y_{ij} = 1|\theta_i) = P_j(\theta_i) = c_i + \frac{1 - c_i}{1 + \exp[-a_j(\theta_i - b_j)]} \tag{3}$$

The $c_i$ parameter allows adding flexibility in the data adjustment since it shapes the probability of an individual with low ability to answer right the item $j$. If an individual has a low ability, the probability of answering right an item of moderate or high difficulty is low. In the case that such individual answers right to an item of these characteristics, it can be because of luck. The three parameters logistic model is one of the most general cases of IRT one-dimensional models: in particular, it generalizes the 1PL and 2PL models. From the expression (3), to obtain a 2PL model it's enough with making $c_j = 0$ and $a_i = K$ to obtain the 1PL.

### 2.1.4 Other one-dimensional models

The previous models are defined as dichotomic as the response measurement has only two possibilities: success or failure. Because of its structure, some items can be adjusted using a model that considers that the choice of a wrong answer can give a partial credit to the task that the item considers to be measuring. A particular case of them is known as the partial credit model (Aitkin & Aitkin 2011). Including this, it exists besides a series of models that relate the individual ability and the probability of answering each one of the multiple choices, among them, the graduated response model or nominal response model. This kind of models is defined as polytomous.

For purposes of this paper, only the 2PL model will be used as a resource for the score of the results of Saber 359 test applied in Colombia in a standardized and periodical way since 2009 to the 5 and 9 grades and to 3 grade since 2012.

For a comprehensive review of all models listed here, the reader may refer to Aitkin & Aitkin (2011) and Hulin et al. (1983).

### 2.1.5 Item parameters estimation

Parameter estimation is conducted via marginal maximum likelihood (Bock & Aitkin 1981, Harwell et al. 1988). The method assumes conditional independence in the answers to different items for individuals with the same score (Kass & Steffey 1989). Let $\mathbf{Y}_i$ be the pattern of responses for the $n$ items of an individual $i$ with $\theta_i$ ability:

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{in})$$

Where $Y_{ij}$ takes the value of 1 if the evaluated $i$ answer the question $j$ correctly and zero when he doesn't. The probability of the evaluated $i$ answer the pattern, given his score $\theta_i$ is expressed as follows:

$$P(\mathbf{Y}_i|\theta_i) = \prod_{j=1}^{n} [P(Y_{ij} = 1|\theta_i)]^{Y_{ij}} [P(Y_{ij} = 0|\theta_i)]^{1-Y_{ij}}. \tag{4}$$

Let $g(\theta)$ be the density function that determines the distribution of the ability of all individuals in the population under study. The marginal probability that any individual of the population obtains a Y response pattern is given by:

$$P(\mathbf{Y}) = \int_{\mathbb{R}} P(\mathbf{Y}|\theta) g(\theta) d\theta \tag{5}$$

This integral can't be treated analytically, but the marginal probability can be approximated through the Gaussian quadrature:

$$\bar{P}_{\mathbf{Y}} \approx \sum_{k=1}^{q} P(\mathbf{Y}|\mathbf{Y}_k) A(\mathbf{Y}_k) \tag{6}$$

Where $q$ is the number of points in the quadrature, $\mathbf{Y}_k$ is a point of the quadrature and $A(\mathbf{Y}_k)$ is a positive weight corresponding to density function $g(\cdot)$. In the method the value of items parameter estimates are chosen such that they maximize the logarithm of marginal likelihood defined by:

$$\log L_M = \sum_{l=1}^{S} r_l \log \bar{P}_{\mathbf{Y}_l} \tag{7}$$

Where $r_l$ is the observed frequency of $\mathbf{Y}_l$ pattern in the sample of students and S is the number of different observed patterns. EM algorithm along the methods Newton-Gauss or Fisher-Scoring are used to solve the implicit equations necessary to find the solution of $\partial \log L_M / \partial \Pi_j$ con $\Pi_j = (a_j, b_j)$.

### 2.1.6 Location and scale of latent distribution

To achieve the estimation of latent distribution is necessary to find its indeterminate measures of location and scale. This indeterminacy arises because in the logit

$$z_j = a_j(\theta - b_j) \tag{8}$$

Any change in the origin of $\theta$ can be controlled by $b_j$ and any change in the unity of $\theta$ can be controlled by $a_j$. To establish the localization parameter, the median of the latent distribution is fixed in 0 and to establish the parameter of scale the variance of latent distribution is fixed in 1.

A convenient way to characterize an arbitrary latent distribution with finite median and variance consist in calculating the density of probability in a finite number of adequate choices of values of $\theta$ and normalizing the densities dividing by the total. These normalized values can be used as the weights $A(\mathbf{Y}_k)$of equaiton (6).

### 2.1.7 Score or ability of individual (test score)

In this section two approaches to fit the score of individuals will be presented, the classical Maximum Likelihood and the Bayesian approach.

**Maximum likelihood estimation**

Maximum likelihood estimation $\hat{\theta}_i$,is the value of that maximizes the following function:

$$\log L(\theta_i) = \sum_{j=1}^{n} \{y_{ij} \log P(Y_{ij} = 1|\theta_i) + (1 - y_{ij}) \log P(Y_{ij} = 0|\theta_i)\}. \tag{9}$$

The equation to be solved is:

$$\frac{\partial \log L(\theta_i)}{\partial \theta_i} = \sum_{j=1}^{n} \frac{y_{ij} - P(Y_{ij} = 1|\theta_i)}{[P(Y_{ij} = 1|\theta_i)][P(Y_{ij} = 0|\theta_i)]} \frac{\partial P(Y_{ij} = 1|\theta_i)}{\partial \theta_i} = 0 \tag{10}$$

Fisher Scoring calculates the maximum likelihood estimator, it depends on Fisher's information which is obtained like that:

$$I(\theta_i) = \sum_{j=1}^{n} a_j^2 P(Y_{ij} = 1|\theta_i)[P(Y_{ij} = 0|\theta_i)]. \tag{11}$$

The solution of the method's iterations is this:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + I^{-1}(\hat{\theta}_t) \left( \frac{\partial \log L(\theta)}{\partial \theta} \bigg|_{\theta=\hat{\theta}_t} \right). \tag{12}$$

The standard error of estimates via maximum likelihood is given by:

$$S.E.(\hat{\theta}) = \sqrt{1/I(\hat{\theta})}. \tag{13}$$

A disadvantage of this method is that there is not a solution to estimate equations when the student answers right or wrong absolutely all the items. These problems don't show up in the other estimate methods.

**Bayesian estimate**

Bayesian estimator is the average of the distribution a posteriori of $\theta$, once response patterns $Y$ has been observed. This can be approximated by Gaussian quadrature as follows:

$$\hat{\theta}_i \cong \frac{\sum_{k=1}^{q} \mathbf{Y}_k P(\mathbf{Y}|Y_k) A(\mathbf{Y}_k)}{\sum_{k=1}^{q} P(\mathbf{Y}|Y_k) A(\mathbf{Y}_k)} \tag{14}$$

This statistic is called expected a posteriori estimator. The measure of its accuracy is the standard deviation a posteriori (DEP), approximated by the following expression:

$$DEP(\hat{\theta}_i) \cong \frac{\sum_{k=1}^{q}(\mathbf{Y}_k - \hat{\theta}_i)^2 P(\mathbf{Y}|\mathbf{Y}_k)A(\mathbf{Y}_k)}{\sum_{k=1}^{q} P(\mathbf{Y}|\mathbf{Y}_k)A(\mathbf{Y}_k)} \tag{15}$$

Weights $A(\mathbf{Y}_k)$ depend on assumed distribution for $\theta$. There are possibilities of having theoretical weights or empirical weights $A^*(\mathbf{Y}_k)$ or subjective weights. This estimator exists for any response pattern and has a lower average error in the population than any other estimator, even the maximum likelihood estimator. It is biased towards the median of the population but the bias is small (Bock & Mislevy 1982).

In Saber 359 test is used the Bayesian estimate. A priori information of parameter $\theta$ is considered normal and the defined number of points of quadrature is 30.

## 2.2 Large-scale assessments in Education

A large-scale evaluation measures what the members of an specific population know and can do with respect to certain skills related to some subjects of interest of any academic program or if they have acquired skills needed for activities in the future. The amplitude of measured subjects in these programs is such that a very large number of content and skills are evaluated (González & Rutkowski 2010). There are many evaluations of this type, but as example of these tests we can mention the Saber 359 test in Colombia, which measures the student's skills from all over the country in language, mathematics, natural sciences and citizen skills in third, fifth and ninth grade, or the PISA (Programme for International Student Assessment) test, which evaluates each three years the skills in mathematics, reading and natural sciences of fifteen years old children.

Large-scale standardized tests usually intend to evaluate an extensive domain of academic content in the population. Because of this, to avoid overloading students and because of economic costs and time, the tests are designed so that a student is given only a fraction of this, it is, a particular combination of items of the test so as to ensure the necessary coverage throughout the population.

This kind of structure is known as multiple matrix sampling or item matrix sampling designs. This approach allows estimating in an accurate way the behavior of skills in the population or in subpopulations and, in turn, the coverage of the whole frame of reference of the evaluation. It allows the reduction of the load of each tested student and the duration of the test in its execution (González & Rutkowski 2010).

Since each student does not answer all the test, the accuracy of individual estimation measurement is sacrificed for the interest to assess the entire frame of reference in the population. Because of this emphasis, the design is not optimum to report individual results, so the obtained results in this kind of evaluations are

always aggregated. For example in Saber 359 test, the minimum grouping level to which results are reported is the school. The maximum, of course, is the national aggregation level. For operational convenience, the items that make up the measuring instrument are assigned to blocks that are then combined in ways according to a particular specification or design. For a comprehensive review of block design, the reader may refer to (González & Rutkowski 2010).

In fifth grade in Saber 359 test, the subject areas assessed are natural sciences, language, and mathematics. To complete measurement of the reference frame 6 blocks were built. Each block has 24 items, it is, 144 items were consolidated for the evaluation. By design, only two blocks of six were supplied to each student that presented the mathematics test. For a comprehensive review of design Saber 359 test, the reader may refer to the technical test report (*Informe técnico SABER 5o. y 9o. 2009* n.d.).

## 2.3 Plausible values multiple imputation

This section shows in a general way what is the method of plausible values method, the score estimation model that it induces and its characterization in mathematical terms. Plausible values method charged on the aggregate results that information which by design cannot be measured in the whole population.

### 2.3.1 Score estimation model

Now, let's assume that we also have auxiliary information $\mathbf{X}$ measurable in each of the individuals in the population. The aim is generally focused on finding the distribution associated with the student's ability as follows:

$$P(\theta|\mathbf{X},\mathbf{Y}) \tag{16}$$

In this case, the distribution of skills in the population is not only conditioned on the response pattern but now it's associated with the values of the exogenous variables of the individual $\mathbf{X}$. The expression (16) can be written as follows:

$$P(\theta|\mathbf{X},\mathbf{Y}) = P(\mathbf{Y}|\mathbf{X},\theta)P(\theta|\mathbf{X}). \tag{17}$$

For this it's necessary to do the following assumptions, known as the conditional independence assumptions:

$\mathbf{Y}$ is conditionally independent of $\mathbf{X}$, it is:

$$P(\mathbf{Y}|\mathbf{X},\theta) = P(\mathbf{Y}|\theta) \tag{18}$$

Elements of vector $\mathbf{Y}$ are conditionally independent, it is:

$$P(\mathbf{Y}|\theta) = \prod_{i=i}^{n} P(Y_i = y_i|\theta) \tag{19}$$

In the case of 2PL model, this expression is the product of probabilities expressed in its functional way as in (3) for $y_i$taking values of 0 and 1.

From the first assumption and by the Bayes theorem following expression is obtained for equation (17):

$$P(\theta|\mathbf{X},\mathbf{Y}) = P(\mathbf{Y}|\theta)P(\theta|\mathbf{X}) \tag{20}$$

### 2.3.2 Latent regression model

In equation (20) is found that an appropriate model for the expression $P(\theta|\mathbf{X})$is missing. For this it's assumed a normal distribution with covariance matrix $\Sigma$ and median given by a linear function of $\mathbf{X}$, it is:

$$P(\Theta|\mathbf{X}) = \Phi(\Theta; \mathbf{X}\Gamma, \Sigma) \tag{21}$$

Whit $\Theta$ the vector of probabilities of the whole population and $\Phi(\cdot)$ the density function probability of a random variable with the normal distribution. This suggests the next regression model:

$$\Theta = \mathbf{X}\Gamma + \epsilon \tag{22}$$

Where $\epsilon \sim N(\mathbf{0}, \Sigma)$ and $\Gamma$ and $\Sigma$are parameters to estimate. The non-observable amount of $\theta$ for each individual in the population depends to some extent on certain variables $\mathbf{X}$ that characterize the population.

**Model parameters estimation $\theta$ EM algorithm**

As the expression (22) is not an ordinary regression model, not any method can be applied to find the estimations of parameters $\Gamma$ and $\Sigma$. Plausible values can be seen as non-observed measurements in a student. In this sense, EM algorithm must be used to find estimations of maximum likelihood.

The first step of the algorithm is formulating the solutions if the vector $\Theta$or response vector is observed. Let $\mathbf{D}$ be the diagonal matrix of the sampling weights of $N$ individuals which induce a sampling design $\mathbf{D} = \text{diag}(w_1, \ldots, w_N)$ and be $p$ the number of covariates observed. The maximum likelihood estimators of parameters are:

$$\hat{\Gamma} = (\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}\Theta \tag{23}$$

and

$$\hat{\Sigma} = \frac{1}{\text{tr}(\mathbf{D})}(\Theta - \mathbf{X}\Gamma)^T\mathbf{D}(\Theta - \mathbf{X}\Gamma) \tag{24}$$

This allows to establish that in the iteration $k$ of estimates algorithm are obtained according to step M as follows:

$$\hat{\Gamma}^{(k+1)} = (\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}\tilde{\Theta}^{(k)} \tag{25}$$

and

$$\hat{\Sigma}^{(k+1)} = \frac{1}{\text{tr}(\mathbf{D})}\left(\text{E}((\Theta - \tilde{\Theta}^{(k)})^T\mathbf{D}(\Theta - \tilde{\Theta}^{(k)})|\mathbf{X},\mathbf{Y},\Gamma^{(k)},\Sigma^{(k)}) + (\Theta^{(k)} - \mathbf{X}\Theta^{(k)})^T\mathbf{D}(\Theta^{(k)} - \mathbf{X}\Theta^{(k)})\right) \tag{26}$$

Step E of iteration k is to find solution to the following a posteriori expectations, necessary to advance in (25) and (26):

$$\tilde{\Theta}^{(k)} = \text{E}(\Theta|\mathbf{X},\mathbf{Y},\Gamma^{(k)},\Sigma^{(k)}) \tag{27}$$

and

$$\text{E}(\Theta^T\mathbf{D}\Theta|\mathbf{X},\mathbf{Y},\Gamma^{(k)},\Sigma^{(k)}) \tag{28}$$

The evaluation of these last two quantities can be performed by two algorithms: integration by quadrature or Laplace information.

To make a comprehensive review of the method, including convergence and the initial values the reader may refer to Rubin (1991).

## 2.4 Valores plausibles

To find estimates of the latent regression's parameters, the EM algorithm can be summarized as follows:

1. To specify or find the initial values $\Gamma^{(0)}$ and $\Sigma^{(0)}$ by estimation via maximum likelihood using quadrature.

2. Step E: To evaluate median and variance of distribution a posteriori of the latent vector $\Theta$ like in expressions (27) and (28). It can be addressed by quadratures or Laplace information, depending on the latent vector's dimension. In one-dimensional models, the integration can be done by quadrature.

3. Step M: Update the values of $\Gamma$ and $\Sigma$ like in expressions (25) and (26)

4. Establish convergence through a suitable criterion. In case it is rejected, return to step 2. In the other case define $\hat{\Gamma}$ y $\hat{\Sigma}$.

As $\hat{\Gamma}$ and $\hat{\Sigma}$ are found, it is possible to define an algorithm of random estimate for R plausible values and for each element of population as follows:

1. Select randomly $\Gamma^{(r)}$ of distribution $N(\hat{\Gamma}, \hat{\Sigma}^{-1}\mathbf{X}^T\mathbf{D}\mathbf{X})$

2. Calculate the mean of assumed distribution for $\theta_i$ as $x_i^T\Gamma^{(r)}$ in $x_i^T$ he row vector of matrix $\mathbf{X}$.

3. To determine a posteriori distribution of $\theta$ shown in the equation (20) using $\hat{\Sigma}$ as the variance matrix of a priori distribution $P(\theta|X)$ of $\theta$ and using besides de response to item model.

4. Select randomly an imputed value $\tilde{\theta}_i^{(r)}$ of a posteriori distribution found in 3 step.

5. Repeat process for $r = 1, 2, \ldots, R$.

Other mechanisms of random selection can be used. In particular, besides this one, the ETS mentions another one in its software DESI (Direct Estimation Software Interactive) as follows:

Calculate the mean of distribution for $\theta_i$ with $x_i^T \hat{\Gamma}$ with $x_i^T$ the row vector of matrix $\mathbf{X}$.

Using $\hat{\Sigma}$ as the variance matrix of a priori distribution $P(\theta|\mathbf{X})$ of $\theta$ shown in the equation (20) using besides the response to item model.

Select randomly an imputed value for $\tilde{\theta}_i^{(r)}$ of a posteriori distribution, found in 2 step.

Repeat the process for $r = 1, 2, \ldots, R$.

In general $R = 5$ is used. The statistical treatment of imputed values is based on producing estimates of parameters of interest in the following via:

Each parameter is estimated for each of the R plausible values and all estimates are averaged.

The standard error of this estimated average is calculated by combining the average sampling error of R estimates with the variance of R estimates.

If the scales on which measurements of the population are projected are well built, the inclusion of any variable of classification in regression shouldn't bias the result of multiple imputations on latent distribution.

# 3 Example: a simulation case

The aim of the simulation is to observe the use of plausible values as an alternative to correct the bias and imprecision caused by the application of just a sample of the items to each individual. To this are considered 4,000 students in 160 schools, and for each of them two associated variables are defined: sex (male or female) related to each student and sector (public or private) related to each school to which it belongs. Table (1) relates the induced averages and deviations for each of the four population groups.

One of the two factors of the simulated population is conditioning (funding school variable) and the other is not (sex). Difference of results between men and women is null, while difference between aggregations of public and private is of 1.414

Table 1: *Mean and standard deviation used to generate the simulation. Source: Own elaboration*

| | Male | Female | Mean |
|---|---|---|---|
| Oficial | -0.707(0.707) | -0.707(0.707) | -0.707(0.707) |
| No oficial | 0.707(0.707) | 0.707(0.707) | 0.707(0.707) |
| Mean | 0.000(0.707) | 0.000(0.707) | 0.000(0.707) |

unities, it is, approximately $\sqrt{2}$ each one of them with a standard deviation of approximately 1..

It's considered that there are 72 items and groups of 12 are randomly selected to conform 6 blocks: A, B, C, D, E y F. A block is a number of questions responding to the same content and difficulty characteristics than the whole test.

Only six booklets (blocks combination) are selected and three simulation scenarios are built like that:

Population stage. All students answer the whole tests, it is, and the 72 items in the 6 blocks are provided to each student.

. Each student answer four blocks of six, it is 48 items (4/6).

. Each student answer two blocks of six, it is 24 items (2/6)

Exercise consists in simulating a set of data for each stage under same conditions: the change of stage 1 to stage 2 consists in removing of observations the two last blocks and at the same time the change of stage 2 to 3 consists in removing of observations the last two blocks. After simulating under normality the score of each student, the values of parameters defining the model is also simulated under normality. The logistical method of two parameters 2PL is used for the exercise. Finally, as an input for estimation phase, it is necessary to simulate the response vector $\mathbf{Y}$.

Let $\tilde{\theta}_i$ be the simulated score for individual i and be $\tilde{a}_j$ and $\tilde{b}_j$ the discrimination and difficulty parameters simulated for the item j. Following steps are used to generate the $Y_{ij}$variable that follows a Bernoulli conditional distribution with $p_{ij} = P(Y_{ij} = 1|\theta_i)$:

1. $u$ is generated with $U \sim U(0, 1)$

2. Calculate

$$\tilde{p}_{ij} = P(Y_{ij} = 1|\tilde{\theta}_i) = \frac{1}{1 + \exp\{-\tilde{a}_j(\tilde{\theta}_i - \tilde{b}_j)\}} \tag{29}$$

3. $u < \tilde{p}_{ij}$ is evaluated. If it's true $Y_{ij} = 1.$ is assigned. If it's false $Y_{ij} = 0$ is assigned.

### Stage 1

In this stage each student presents all the blocks. In this case it is not necessary to impute his average score according to some aggregate, it is, it is not necessary to use plausible values and the purpose of test will be different. Design can be shown as in table 2:

Table 2: *Booklets design in stage 1. Source: own elaboration*

| Booklet | Bloq. 1 | Bloq. 2 | Bloq. 3 | Bloq. 4 | Bloq. 5 | Bloq. 6 |
|---|---|---|---|---|---|---|
| 1 | A | B | C | D | E | F |
| 2 | F | A | B | C | D | E |
| 3 | E | F | A | B | C | D |
| 4 | D | E | F | A | B | C |
| 5 | C | D | E | F | A | B |
| 6 | B | C | D | E | F | A |

This stage assumes that the entire population of students is measured in each of the items of the test and serves as a reference to compare results of multiple imputations when each student doesn't present all the items. Using the Y vector, the model is adjusted for each parameter and it's compared against induced values in the simulation. Figure 1 shows on the left, the simulated difficulty for each of the 72 items against the estimated difficulty and, on the right, the distribution of the 72 items according to the simulated or estimated difficulty.

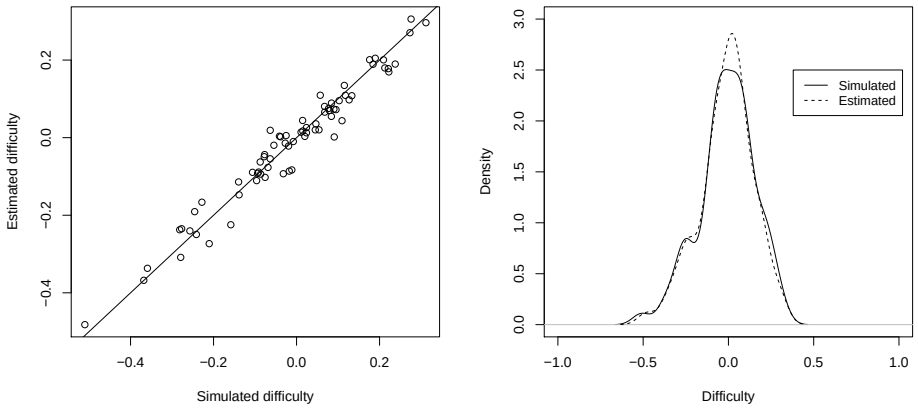


Figure 1: *a). Plot of estimated and simulated difficulty / b).Distribution of simulated and estimated difficulty in the 72 items. Source: own elaboration.*

Figure 1 shows that fit of parameters model's difficulty is good and that, except for the usual variability, the association between simulated and observed is evident. Figure 2 shows on the left, for each of the 4000 individuals, the simulated score

versus the estimated score and, on the right, the distribution of the 4000 individuals according to their simulated or estimated score.

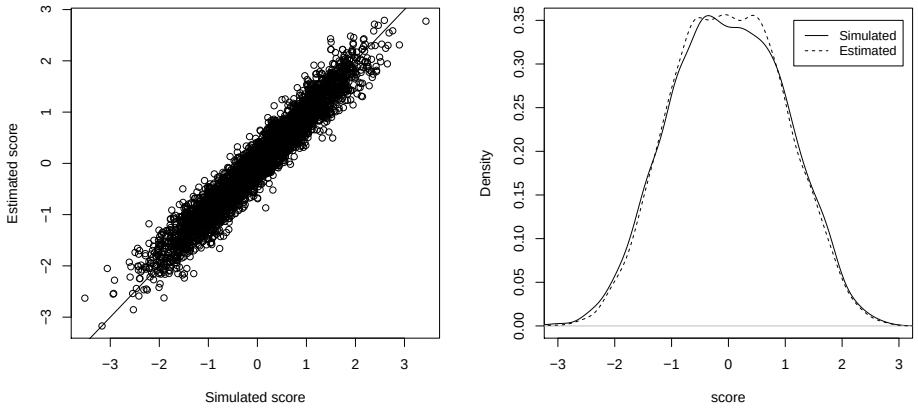


Figure 2: *a). Plot of simulated and estimated score b). Distribution of simulated and estimated score in the 4000 individuals. Source: own elaboration.*

The fit of each individual's score by the model is good and the association between simulated and observed is important. Figure 3 relates the density of the skill's distribution of the population in general and when added by the variable of the sector of the schools. The estimator used is the expected score a posteriori.

The public schools show an average score lower than the private. These amounts are -0.6651 and 0.6674 respectively, with a standard deviation of 0.708 y 0.710. The sex variable is not significant since the average between men and women obtain an average score of -0.003 and 0.004. Magnitudes are very close to what induced by simulation, but not exact, so this result is the reference point of what should be measured when information is no longer observed, as in the case of the following two stages.

### Stage 2: Design 4/6 blocks

In this stage each student answers four blocks of 6. Table 3 shows the design.

Table 3 shows that 4 of 6 students answer each block, it is the 2/3 of the population. Under this scenery arises the problem of incomplete information of the test in each student. Figure 4 shows on the left, the simulated difficulty for each of the 72 items against the estimated difficulty with the characteristic that it doesn't use the entire population; on the right, it shows the distribution of the 72 items according to the simulated and estimated difficulty under such conditions.

Figure 4 shows that the fit of the model's difficulty parameter is good, although the association between the simulated and the observed gets distorted compared to the observed in the figure of the first stage 1. In the case where less information

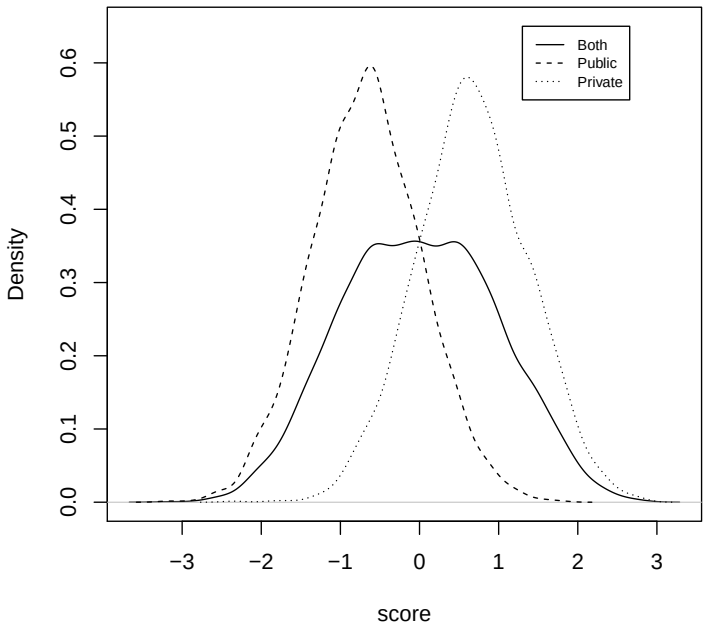


Figure 3: *Score distribution in the population and by funding school variable. Source: own elaboration.*

Table 3: *Booklets design in the stage 2. Source: own elaboration.*

| Booklet | Block. 1 | Block. 2 | Block. 3 | Block. 4 |
|---|---|---|---|---|
| 1 | A | B | C | D |
| 2 | F | A | B | C |
| 3 | E | F | A | B |
| 4 | D | E | F | A |
| 5 | C | D | E | F |
| 6 | B | C | D | E |

than expected is observed, estimates of the model show a loss of accuracy and increase the risk of including bias. Figure 5 shows on the left, the simulated score of each of the 4000 individuals against the estimated under this scenery and on the right, the distribution of the 4000 individuals according to their simulated and estimated score.

Figures show that the fit of each individual's score by the model is good although the association between the simulated and the observed acquires a remarkable

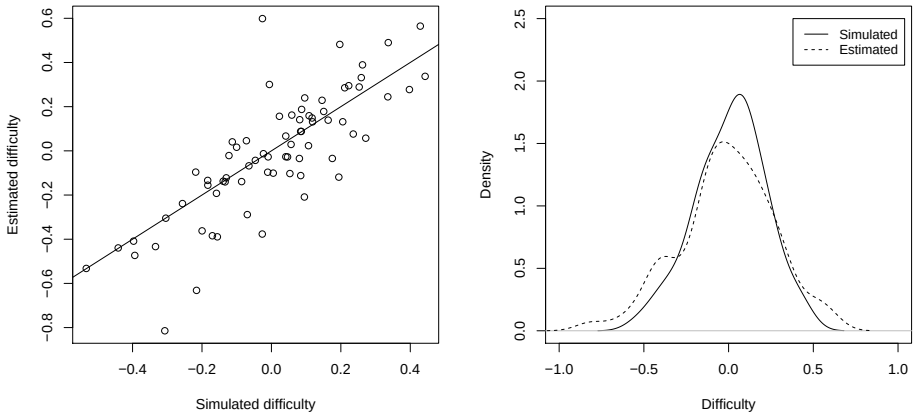


Figure 4: *a). Plot of simulated and estimated difficulty (4/6) / b). Distribution of simulated and estimated difficulty (4/6) in the 72 items. Source: own elaboration*

variability. When losing information in each student, the score's variance is underestimated. One of the direct implications of this type of phenomenon is that the type I error can be inflated, eventually causing the declaration of differences in some subpopulation groups when such differences do not really exist. Figure 6 shows the density of the population's abilities distribution in general and by funding variable of schools.

The public schools show an average lower than the private, keeping the magnitude relation with which the simulation was induced. However, these amounts are $-0.4573$ and 0.4532 respectively, with a standard deviation of 0.665 and 0.664. Like in the previous case, we found that the sex variable is not significant since on average men and women obtain a score of $-0.008$ y 0.008, respectively. Relations are very close to the induced by simulation and the observed in the first stage, but magnitudes show important changes. This is because the tests are not wholly supplied to the students and, each one of them answers only one part of the same. So, it makes evident that aggregate of expected score a posteriori is a biased estimator of average.

**Stage 3:Design 2/6 blocks**

In this scenery, each student answers 2 blocks of 6. Table (4)shows the design.

Table(4) shows that 2 of 6 students answer each block, it is 1/3 of the population. Figure 7 relates on the left side the simulated difficulty of each of the 72 items against the estimated difficulty, and on the right side, it shows the distribution of the 72 items according to the simulated and estimated difficulty under such conditions.

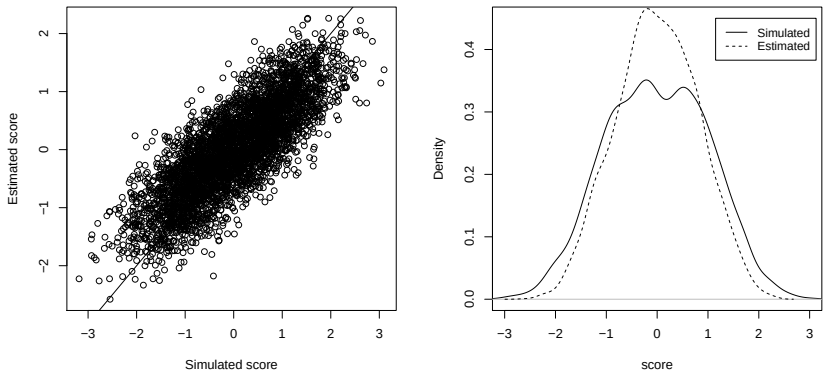


Figure 5: *a). Plot of simulated and estimated score (4/6). b). Distribution of simulated and estimated score (4/6) in the 4000 individuals. Source: own elaboration.*

Table 4: *Design of booklet in the stage 3. Source: own elaboration.*

| Cuadernillo | Bloq. 1 | Bloq. 2 |
|---|---|---|
| 1 | A | B |
| 2 | F | A |
| 3 | E | F |
| 4 | D | E |
| 5 | C | D |
| 6 | B | C |

The figure shows that the fit of the difficulty parameter of the model is good, although the association between the simulated and the observed gets more distorted than the immediately previous stage. Figure 8 shows on the left side the simulated score for each of the 4000 individuals according to their simulated and estimated score.

When losing information in each individual, score's variance is underestimated. Figure 9 shows the density of the abilities distribution and when it is aggregated by the variable funding level school.

Public schools show an average lower that the private, keeping the magnitudes relation with which simulation was induced and which has been observed in both stages mentioned before. Magnitudes in this scenery are -0.3690 and 0.3476 respectively, with the standard deviation of 0.633 y 0.631. On average, men and women obtain a score of -0.005 y 0.005 respectively. With respect to previous scenery, it's easy to deduce that the aggregate result in this one, tends to minimize differences between aggregate averages of public and privates schools.

**Stage 4: Plausible values for a design 2/6 blocks**

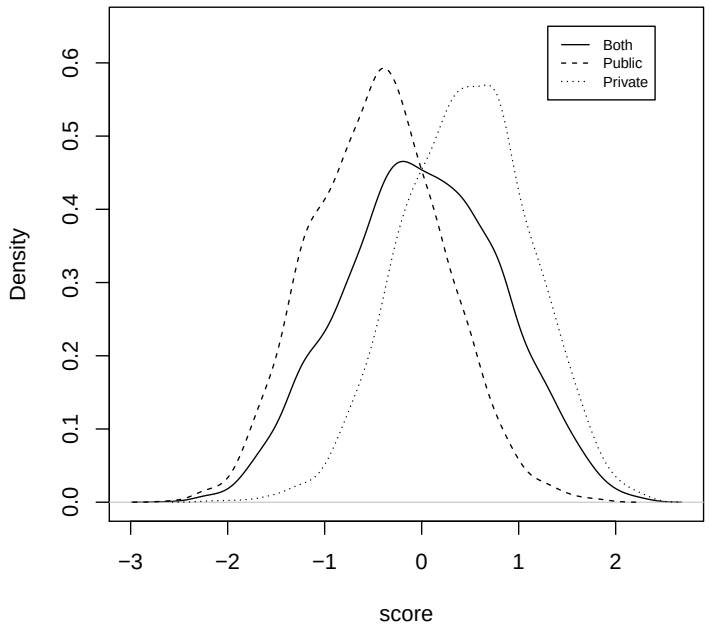


Figure 6: *Score's distribution in population and by funding variable school. Source: own elaboration*

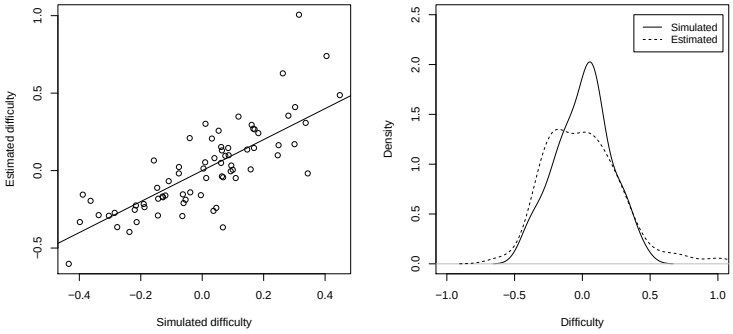


Figure 7: *a). Plot of simulated and estimated difficulty (2/6). b). Distribution of simulated and estimated difficulty (2/6) in the 72 items. Source: own elaboration.*

It is observed that as the number of evaluated items by individual decrease, the

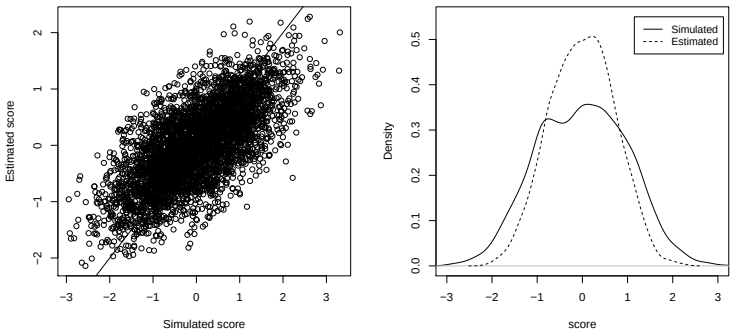


Figure 8: *a). Plot of simulated and estimated score (2/6) /b). Distribution of simulated and estimated score (2/6) in the 4000 individuals. Source: own elaboration.*

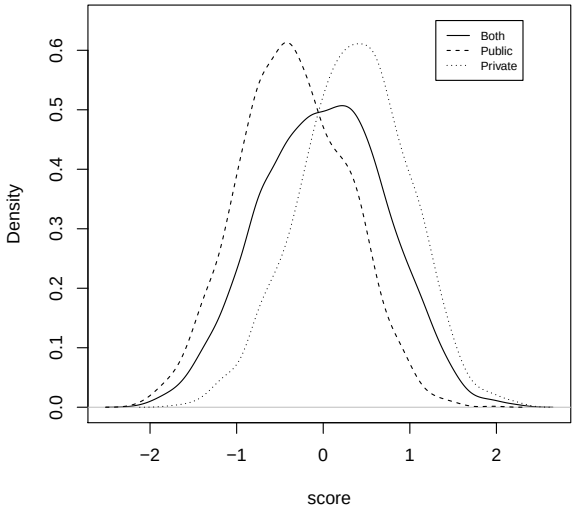


Figure 9: *a). Plot of simulated and estimated score (2/6). b). Distribution of simulated and estimated score (2/6) in the 4000 individuals. Source: own elaboration.*

aggregate results according to some variables of interest suffer a transcendental change that doesn't correspond to reality. Figure (10) shows the change of distribution of estimated abilities of all individuals belonging to public schools.

The figure shows an important bias to the right. This is also because of the used

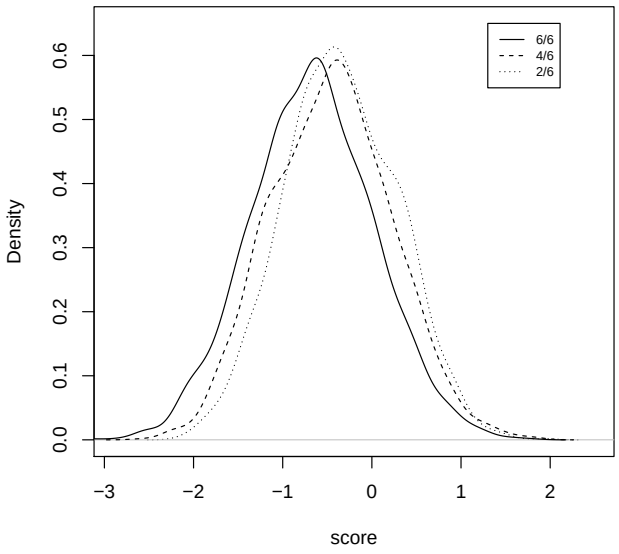


Figure 10: *Score's distribution in the population and by funding school variable. Source: own elaboration.*

estimator is slightly biased toward the average of the population which is 0. To correct this it's necessary to make the multiple imputation procedures. In this exercise, five plausible values of the mentioned a posteriori distribution in (17) are chosen for each individual and statistics of interest are calculated like in the section 2.3. Figure 11 associates the change of the score's distribution in this group in the three mentioned stages along with one of the plausible values.

It is observed that in the stage of plausible values, distribution gets closer to the first stage, correcting the bias and variability above all that loses in the scenery of the two applied blocks of six. Means are shown in tables 5 and 6.

Table 5: *Mean and deviation by funding variable school and sex in the theoretical and first stage. Source: own elaboration.*

| Sex | Funding | Theoretical stage | First stage (6/6) |
|---|---|---|---|
| M | Public | -0.71 (0.71) | -0.6651 (0.6922) |
| M | Private | 0.71 (0.71) | 0.6585 (0.7324) |
| F | Public | -0.71 (0.71) | -0.6672 (0.69) |
| F | Private | 0.71 (0.71) | 0.6764 (0.6787) |

With the table can be verified that it exists observed variability loss when not the entire items are applied to the whole population. However, the method of

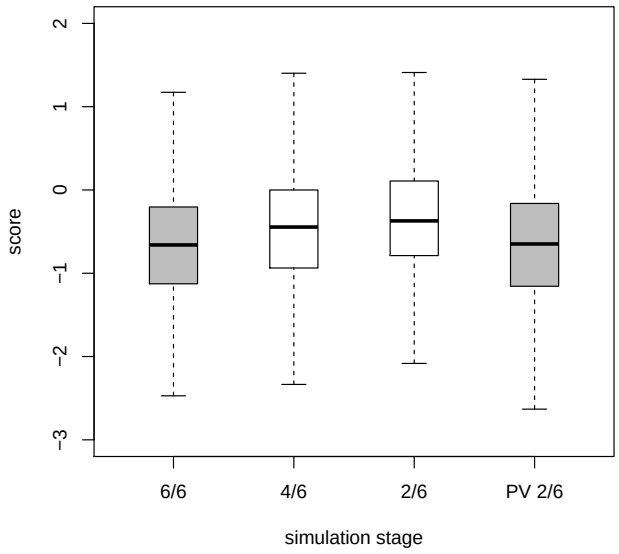


Figure 11: *Score's distribution in the public schools in the four stages. Source: own elaboration.*

Table 6: *Mean and deviation by funding variable and sex in the (4/6) and (2/6) stages. Source: own elaboration.*

| Sex | Funding | Second stage (4/6)) | Third stage (4/6) | Fourth stage (2/6 - PV) |
|---|---|---|---|---|
| M | Public | -0.4573 (0.6653) | -0.369 (0.6338) | -0.6842 (0.7411) |
| M | Private | 0.4396 (0.658) | 0.3575 (0.6105) | -0.6226 (0.7541) |
| F | Public | -0.4494 (0.6857) | -0.3266 (0.6279) | 0.6861 (0.7466) |
| F | Private | 0.4667 (0.6706) | 0.3377 (0.6508) | 0.6252 (0.7514) |

plausible values helps to correct this variability and helps in this specific case to the bias correction.

# 4 Conclusions

Large-scale assessment uses random blocks to guarantee the necessary coverage imposed by the defined reference frame and to guarantee the economy and practicality of the test in this field. This implies a considerable loss of information that should be treated carefully. The non-appropriate treatment of this problem could have serious consequences in the aggregate statistics and in the choice making with

respect to the population under study.

The estimator used in this exercise to calculate the score of each individual is slightly biased toward the population mean. The exercise shows that such a bias increases as the total of blocks measured in each individual decreases. Besides losing information that must be observed, variability over the results of the distribution of the latent feature in the population decreases. This can cause wrong inferences, increasing the possibility of making type I error.

The method of plausible values is a specific case of data imputation through the inclusion of classification's variables of individuals to generate a latent distribution a posteriori. The method helps to control the loss of mentioned variability and, in the particular case of this estimator, to correct the obtained bias.

**Received: February 22, 2016**
**Accepted: April 4, 2016**

## References

Aitkin, M. & Aitkin, I. (2011), *Statistical Modeling of the National Assessment of Educational Progress*, Nueva York: Springer.

Bock, R. D. & Aitkin, M. (1981), 'Marginal maximum likelihood estimation of item parameters: Application of an em algorithm', *Psychometrika* **46**(4), 443–459.

Bock, R. D. & Mislevy, R. J. (1982), 'Adaptive EAP estimation of ability in a microcomputer environment', *Applied psychological measurement* **6**(4), 431–444.

González, E. & Rutkowski, L. (2010), 'Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments', *IERI monograph series: Issues and methodologies in large-scale assessments* .

Harwell, M. R., Baker, F. B. & Zwarts, M. (1988), 'Item parameter estimation via marginal maximum likelihood and an em algorithm: A didactic.', *Journal of Educational and Behavioral Statistics* .

Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983), *Item response theory: Application to psychological measurement.*, Dow Jones-Irwin, Homewood, IL.

*Informe técnico SABER 5o. y 9o. 2009* (n.d.), Technical report.

Kass, R. & Steffey, D. (1989), 'Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models)', *Journal of the American Statistical Association* **84**(407), 717–726.

Rubin, D. B. (1991), 'EM and beyond', *Psychometrika* **56**(2), 241–254.